International Journal of
Geo-Information

MDPI

Article

Influence of Sample Size on Automatic Positional Accuracy Assessment Methods for Urban Areas

Francisco J. Ariza-López, Juan J. Ruiz-Lendínez * and Manuel A. Ureña-Cámara

Department of Cartographic, Geodetic Engineering and Photogrammetry, University of Jaén, 23071 Jaén, Spain; fjariza@ujaen.es (F.J.A.-L.); maurena@ujaen.es (M.A.U.-C)

* Correspondence: lendinez@ujaen.es; Tel.: +34-953-212-470

Received: 7 May 2018; Accepted: 27 May 2018; Published: 28 May 2018

Abstract: In recent years, new approaches aimed to increase the automation level of positional accuracy assessment processes for spatial data have been developed. However, in such cases, an aspect as significant as sample size has not yet been addressed. In this paper, we study the influence of sample size when estimating the planimetric positional accuracy of urban databases by means of an automatic assessment using polygon-based methodology. Our study is based on a simulation process, which extracts pairs of homologous polygons from the assessed and reference data sources and applies two buffer-based methods. The parameter used for determining the different sizes (which range from 5 km up to 100 km) has been the length of the polygons' perimeter, and for each sample size 1000 simulations were run. After completing the simulation process, the comparisons between the estimated distribution functions for each sample and population distribution function were carried out by means of the Kolmogorov–Smirnov test. Results show a significant reduction in the variability of estimations when sample size increased from 5 km to 100 km.

Keywords: accuracy; sample size; simulation; matching; automation

1. Introduction

The assessment of the positional accuracy of cartographic products has always been of great importance. However, nowadays it is a matter of renewed interest because of the need for greater spatial interoperability, supporting new ways of mapping based on Volunteered Geographic Information (VGI) and Spatial Data Infrastructures (SDI). The products derived from these new ways of mapping require the integration of spatial data (inputs) from sources with heterogeneous levels of quality. These levels of quality must be well-understood, not only in order to develop the integration process successfully, but also to determine whether the final quality of the output data fits the users´ requirements. This is why it is necessary to implement more efficient accuracy assessment procedures, which give us a fast and easy evaluation of the quality of these new cartographic products. From our point of view, this can only be achieved by increasing the levels of automation of such procedures.

Traditionally, positional accuracy has been evaluated by means of the positional discrepancies between the apparent location of a spatial entity recorded in a Geospatial Data Base (GDB) and its true (real world) location. However, the task of identifying the true location of a spatial entity by means of topographic field surveying is often not technically or economically feasible (although this will largely depend on the size and accessibility of the geographical area to be evaluated, see as an example [1]). This all decreases the final efficiency of the accuracy assessment processes. In order to overcome such inconveniences, positional accuracy could also be defined by measuring the differences between the location of a spatial entity stored in a GDB (tested or assessed data source) and its location determined by another GDB (reference data source) of higher accuracy. Thus, if the accuracy of the second GDB is high enough, then the unmeasured difference between its information and the real world location can be ignored. This way of defining positional accuracy, proposed by Goodchild and Hunter [2], has inspired the development of new approaches aimed at increasing the automation level of the accuracy assessment procedures. Such approaches are based on spatial data matching mechanisms, which thus acquire a determining role in automatically identifying homologous spatial entities between the two data sources (tested and reference).

Spatial data matching is a relevant research field in Geographic Information Sciences, with many direct and indirect applications. Among them, we highlight data conflation and data quality evaluation. The term conflation is used to describe procedures related to combining geographical information of several scales and precisions, transferring attributes from one dataset to another or adding missing features [3], with the main goal of obtaining an enriched product that is "better" than the previous two [4–7]. Overall, conflation procedures are commonly used in computer science and remote sensing fields [8–11], and mainly in the cartographic updating of urban areas [12–14].

On the other hand, and following Xavier et al. [15], spatial data matching can also be used in data quality assessment approaches, such as inconsistency detection [16,17], positional accuracy [18–21], completeness [22], and thematic accuracy [23]. In the case of positional accuracy, these same authors propose the development of a web service for quality control, whose key process is a simple matching process [24–26]. Overall, it could be said that all the matching mechanisms used in quality assessment approaches share a common characteristic: all of them use objective measures for evaluating the degree of similarity between two GDBs. These measures can be classified according to the nature of the measured quantity: geometry, topology, attributes, context, and semantics [15]. In the specific case of the matching mechanisms applied to positional accuracy assessment, similarity measures employed to match elements are related to the geometric properties of spatial features [21].

On the basis of the above, in our previous studies [19,21] we proposed a matching-based methodology for automating the positional accuracy assessment of a GDB, by using another GDB as reference source and polygonal features as spatial entities on which to determine the degree of similarity (or dissimilarity) between both data sets. Polygons are closed by lines, so their positional behavior can be analyzed through their boundaries using buffer-based methods. Specifically, we used buildings from urban areas, because they are a huge set of polygonal features in any GDB, and therefore the positional assessment derived from them will be deemed statistically significant. In addition, they have a wide spatial distribution in cartography and more temporal permanence than other polygonal features. The proposed matching mechanism determined a set of homologous polygons between both GDBs, using a weighted combination of geometric measures. Thus, the assigned weights among measures were calculated from a supervised training process using a genetic algorithm (GA) [27] (this process was externally evaluated by fifteen reviewers selected from a pool of internationally recognized experts who certified its robustness and quality—see [21]). Specifically, the geometric measures employed quantified the absolute location of the polygons by means of their overlapping areas and geometric properties, such as the length of the perimeter and the area of a polygon. In addition, some shape measures were employed for assessing the geometric form of polygons, such as the number of concave and convex angles, moment of inertia, and the area of the region below the turning function. This mechanism also produced a Match Accuracy Value (MAV) (see [19,21]). The MAV was obtained as a linear combination of the geometric measures computed for two homologous polygons, and the weights resulting from the supervised training phase. Such an indicator was relevant for our work, since the setting of a threshold value for it (by means of a confusion matrix and with a certain confidence level) allowed us to select only 1:1 corresponding polygon pairs among all the possible correspondences, thus avoiding the acceptance of both erroneously-matched polygons (false positive) that appear in the cases of 1:n or n:m correspondences (multiple matching cases often associated with generalization processes) and unpaired polygons (null matching cases derived from completeness or updating problems). Figure 1 illustrates all the possible correspondences between polygons after applying our matching mechanism.

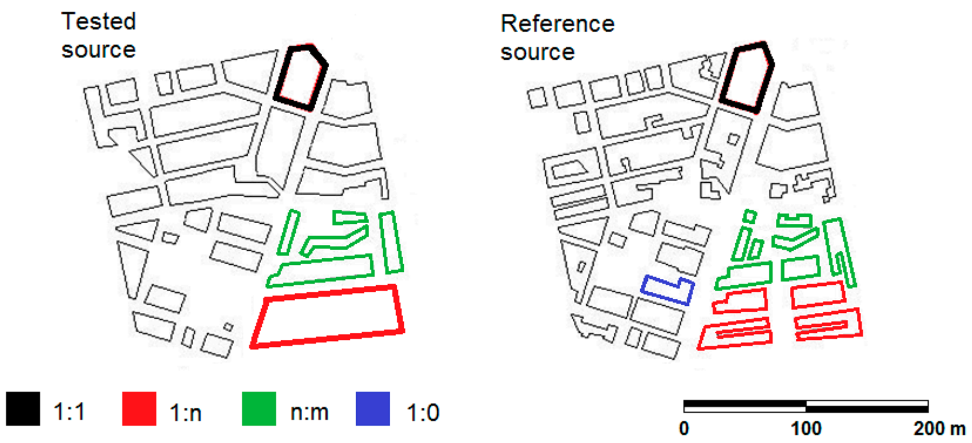

Figure 1. Classification according to the matching case. 1:1 correspondences (highlighted in black), 1:n correspondences (highlighted in red), n:m correspondences (highlighted in green), and unpaired polygons (highlighted in blue).

Having obtained the homologous polygons (1:1 corresponding polygons pairs), we used two positional accuracy assessment methods, based on buffer generation on the polygons' perimeter lines: he simple buffer overlay method (SBOM) [2] and the double buffer overlay method (DBOM, originally the buffer statistics overlay method developed by Tveite and Langaas [28]). These methods allowed us to compute the displacement between two polygonal features and represent two basic and different cases because of the different relationships in the assessment: the first is a line–buffer-based option, and the second is a buffer–buffer-based option. Finally, the results obtained by applying the methods described above demonstrated the viability of the proposed approach, because they confirmed the results obtained by means of the traditional positional accuracy assessment procedures, using GPS data acquisition applied to the same geographical area [19].

However, despite all this, there are still important aspects with relation to the Automatic Positional Accuracy Assessment (APAA) of urban GDBs that have not yet been determined or addressed. Such is the case with sample size and sample distribution. Specifically, in the case of the two buffer-based methods mentioned above (SBOM and DBOM), there are no recommendations about adequate sample size. In addition, the previous studies dealing with positional accuracy assessment by means of such methods are very scarce, and employ only linear elements, such as roads or coast lines, as control elements [2,28–33]. Thus, after identifying and extracting these elements from both GDBs (tested data source and reference data source), they are manually edited and matched, so the level of automation is null. Table 1 summarizes some of these previous studies (sorted by publication date).

Table 1. Review of some previous studies dealing with positional accuracy assessment by means of buffer-based methods.

Characteristics	Method	Sample Size (km)	Tested Scale	Reference Scale
Abbas et al. [29]	SBOM	266	1:100,000	1:25,000
Goodchild and Hunter [2]	SBOM	247	1:1,000,000	1:25,000
Kawaga et al. [30]	SBOM	0.76	1:2500	1:500
Tveite and Langaas [28]	DBOM	-	1:1,000,000	1:250,000
Johnston et al. [31]	SBOM	-	1:24,000	0.7 m (Image)
Van Niel and McVicar [32]	SBOM	466	1:50,000	1.5 m (GPS)
Mozas [33]	SBOM/DBOM	136	1:25,000	2 m (GPS)

As shown in the table, most of these authors provide the sample size employed in their work. However, none of them explain how that parameter was determined. Therefore, as in these previous studies, we consider it important to establish specific criteria as well as guidance in order to define sample size when APAA methods are employed for assessing the positional accuracy of GDBs, because this parameter might influence the uncertainty of the estimated values.

The objective of this paper is to analyse the influence of sample size in terms of uncertainty when estimating the planimetric positional accuracy of urban data (buildings) belonging to territorial GDBs, by means of APAA methods based on buffered polygons (SBOM and DBOM).

The rest of the paper is divided into the following main sections: the next section presents the buffer-based methods applied and their adaptation to the line-closed case. Then the two urban GDBs used are presented, together with the positional accuracy estimation obtained from them. The following section explains the simulation process applied in order to estimate the positional accuracy for different sample sizes. Results are presented and discussed in the last section. Finally, general conclusions are presented.

2. The Buffer-Based Methods Used and Their Adaptation to the Line-Closed Case

Euclidean distance, or the Euclidean metric, is the typical measure of positional accuracy when point-to-point relations between two spatial data sets are used. In this sense, APAA methods are not an exception, as demonstrated by Ruiz-Lendínez et al. [20]. Thus, after automatically identifying homologous points between previously-matched polygons (using a metric for comparing their turning functions), these authors apply a standard based on the Euclidean distance between points to assess the positional accuracy of the tested data source. However, Euclidean distance encounters significant difficulties when applied to linear elements to determine their relative positional accuracy, because only when these elements are parallel does this measure achieve a complete meaning. Therefore, the distance between non-point features is a difficult concept; although there are several proposals developed for positional accuracy assessment using linear elements based on distance measurements (the Hausdorff distance method [29,34] and the mean distance method [35–37]), when using polygons their positional behavior and geometrical similarity can be more accurately and efficiently analyzed through their boundaries, using buffer-based methods (because of the lack of the above-mentioned difficulties). Specifically, and as already mentioned, the two buffer-based methods selected and their base references are the SBOM [2], and the DBOM [28,38]. According to [2,28,38], both methods give a quantitative assessment of the geometric accuracy of a line relative to another line of higher accuracy. In addition, they are iterative, and both the size of the first buffer w_0 and the value by which it is increased Δw (step size) must be set on the basis of the spatial accuracy of the reference GDB (whose value is, in principle, well-known), and the approximate spatial accuracy of the tested GDB (this value can be estimated on the basis of information provided by the producer [28,38]). The assignment of adequate values to the aforementioned parameters (w_0 and Δw) will allow us to achieve a fast stabilization of the distribution function, which acts as a signature of the tested GDB (Figures 2c and 3c). Finally, it must be noted that the value of Δw may change depending on the level of detail required. Therefore, at the beginning of the buffer operation (when more detail is usually required), Δw usually takes smaller values than at the end of it (when coarser steps are used).

2.1. The Single Buffer Overlay Method (SBOM)

Based on buffer generation on the line of the source of greater accuracy (Q), this method determines the percentage of the controlled line X that is within this buffer (Figure 2a). By increasing the width w of the buffer, we obtain a probability distribution of inclusion of the controlled line inside the buffer of the source of greater accuracy. The same can be done with all the linear entities in a control sample or a complete GDB, obtaining an aggregated distribution curve that shows a distribution function of the uncertainty of each database for several levels of confidence.

Originally proposed by Goodchild and Hunter [2], this method has been applied in a wide range of studies, such as the evaluation of road networks [39], accuracy assessment of conflation of raster maps and orthoimagery [40], and positional accuracy control of GDBs [32]. This method overestimates the error, because the measures of the error distances are perpendicular to the buffered line; however, the displacement can actually be more complex. In our case, this method had to be adapted to a line-closed case (polygons) (Figure 2b). This adaptation was essentially based on buffer generation on the perimeter of the polygon belonging to the source of greater accuracy *Q*. After this, the percentage of the perimeter of the controlled polygon *X* within this buffer was computed. Finally, as in the case of linear entities, we were able to obtain an aggregate distribution function of the uncertainty for several levels of confidence when all polygons from a sample were used.

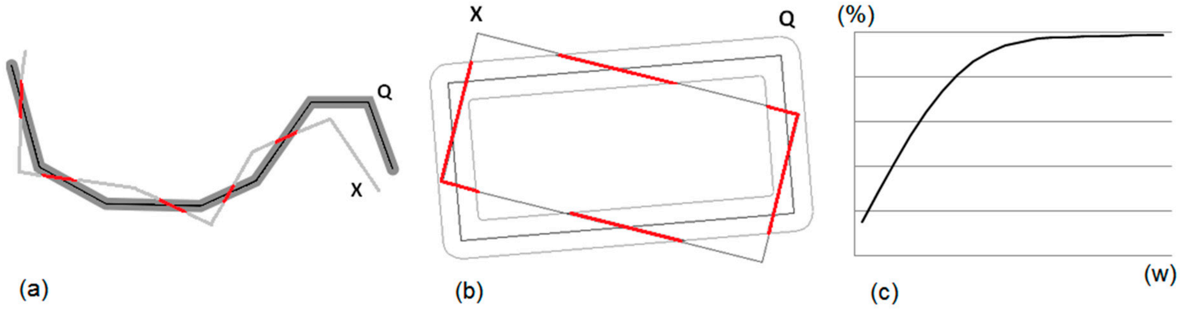

Figure 2. (**a**) The single buffer overlay method (SBOM); (**b**) its adaptation to line-closed case (polygons); and (**c**) its empirical probabilistic distribution function.

2.2. The Double Buffer Overlay Method (DBOM)

Originally proposed by Tveite and Langaas [28], this consists of the generation of buffers (with a width of w) around the two lines—X from the tested source and Q from the source of greater accuracy, denoted as XB and QB, respectively—and analyses the situations that arise when buffers intersect in space (Figure 3a). Thus, four different types of areas result from the buffer and overlay operations: areas that are inside both the buffer of X (XB) and the buffer of Q (QB) ($XB \cap QB$, also denoted as common region), areas that are inside XB and outside QB ($XB \cap \overline{QB}$), areas that are outside XB and inside QB ($\overline{XB} \cap QB$), and finally areas that are outside XB and outside QB ($\overline{XB} \cap \overline{QB}$, also denoted as an outer region).

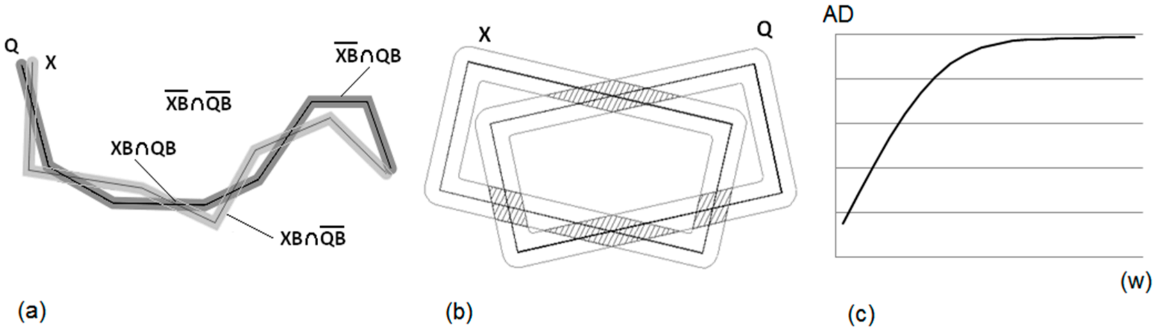

Figure 3. (**a**) The double buffer overlay method (DBOM); (**b**) its adaptation to the line-closed case (polygons); and (**c**) its average displacement estimation function.

Although the area ($XB \cap QB$) compared to the total area from XB or XQ could be used as a good measure of accuracy, we have used another indicator proposed by the above authors for evaluating the deviation of line X from line Q: the average displacement (AD) for a buffer width w (Equation (1)). This measure estimates displacement or similarity by using the area inside QB of the X line, because for similar lines the area ($XB \cap QB$) predominate, but as the lines become more different, the area of the other two will increase as a function of the size of the displacements.

$$AD = \pi w \frac{Area\ (\overline{XB} \cap QB)}{Area\ XB} \tag{1}$$

In their study, Tveite and Langaas applied this method using the 1:250,000 National Map Series of Norway as the reference dataset to assess the digital chart of the world at a 1:1,000,000 scale produced by the Defense Mapping Agency (DMA), United States. As in the case of SBOM, this method had to be adapted to the line-closed case (polygons) (Figure 3b). This adaptation consisted of the generation of buffers around the perimeter lines of the two polygons (X and Q) and the subsequent calculation of the average displacement.

3. The Two Urban Geospatial Databases

As has already been stated, this study takes as its starting point our previous work [19], in which one official GDB was assessed by means of another one of higher accuracy. This section presents the constraints to which these two GDBs were subjected, and a general description of each of them.

With regard to the first aspect, there were three basic conditions that needed to be fulfilled by the two GDBs in order to apply our APAA procedure. These conditions or constraints, to which any other GDB under evaluation should be subject, are included in what have been termed the acceptance criteria:

1. Coexistence criterion (CC): all the elements (in our case, buildings represented by means of polygons) used to apply our APAA procedure must exist in both GDBs. This basic principle, which seems obvious, is not always fulfilled in the real world, since two GDBs generated at different scales are normally not at the same generalisation level [28].
2. Independence criterion (IC): the two GDBs must be independently produced, and in turn, neither of them can be derived from another cartographic product of a larger scale through any process, such as generalisation, which means that their quality has not been degraded.
3. Interoperability criterion (IOC): It is necessary to ensure the interoperability between both GDBs according to the following aspects:

    a. Geometric interoperability: In this case, it is defined in purely cartographic terms, so it will hereinafter be referred to as *cartographic interoperability*. Two GDBs occupying the same geographic region must be comparable, both in terms of reference system and cartographic projection.
    b. Semantic interoperability: there must be no semantic heterogeneity between both GDBs—that is, differences in intended meaning of terms in specific contexts [15]. In this aspect, interoperability must occur at two different levels: schema and feature.
    c. Topological interoperability: the topological relationships must be preserved. In this sense, it can be stated that topological interoperability is a consequence and hence the basis of the two previously described interoperability processes (geometric and semantic) [7].

With regard to the description of both cartographic products, the GDBs used were two official cartographic databases in Andalusia (Southern Spain). Specifically, as the tested source we used the BCN25 (“Base Cartográfica Numérica E25k”) and as the reference source we used the MTA10 (“Mapa Topográfico de Andalucía E10k”).

The MTA10 is produced by the Institute of Statistics and Cartography of Andalusia (Spain) and referenced to the ED50 datum. The MTA10 is a topographic vector database with complete coverage of the regional territory, and is considered to be the official map of Andalusia. It is composed of 2750 sheets obtained by manual photogrammetric restitution. This product includes a vector layer of buildings (city blocks), which contain a sufficient quantity of geometrical information to be able to compute both the shape and geometric measures employed for assessing the geometric form of polygons.

This dataset was used as the reference source because of its higher, a priori, positional accuracy. Thus, its declared positional accuracy is RMSE = 3 m [41].

The BCN25 is produced by the National Geographic Institute of Spain [42] and references the European Terrestrial Reference System 1989 ETRS89 datum. The dataset of the BCN25 is composed of 4123 sheets of 5′ of latitude by 10′ of longitude, which cover the whole national territory of Spain. In addition, and just as in the previous case, the map is presented as a set of vector covers distributed by layers, including a vector layer of buildings (city-blocks) that contains the same type of geometrical information as the MTA10, thus allowing us to determine the degree of similarity between both data sets at polygons level. Finally, the BCN25's planimetric accuracy has been estimated at around 7.5 m, although this varies depending on the type of entity considered [42]. Therefore, the BCN25 was used as the tested source.

The urban areas selected were included in three sheets of the MTN50k (National Topographic Map of Spain, at scale 1:50,000) (Figure 4a). Figure 4b shows two examples of polygonal features corresponding to buildings belonging to MTA10 and BCN25.

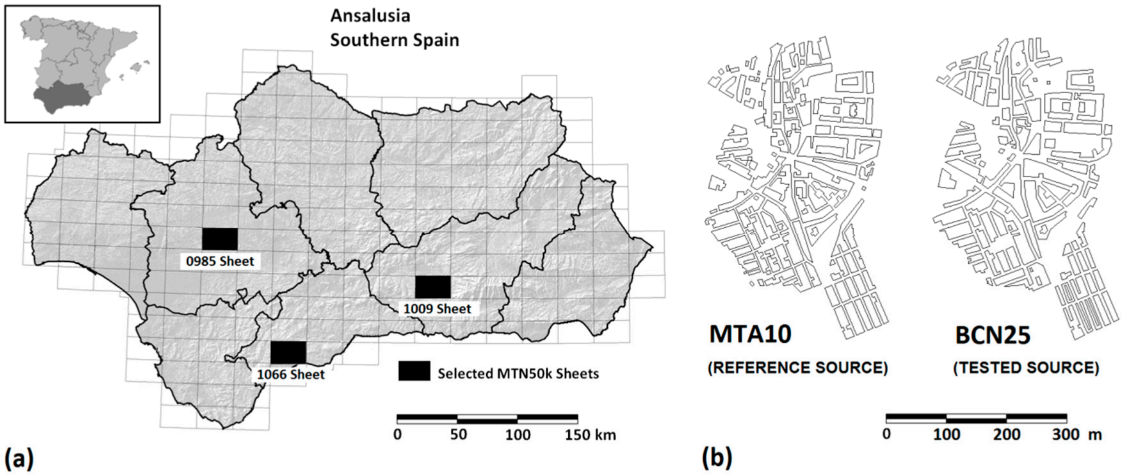

Figure 4. (**a**) Selected sheets of the MTN50K of the region of Andalusia (Spain) and its administrative boundaries; and (**b**) examples of polygonal features belonging to MTA10 and BCN25.

Characterization of the BCN25 Positional Accuracy by Means of the Single Buffer Overlay Method and the Double Buffer Overlay Method

This subsection presents the degree of fulfilment of the constraints to which our GDBs were subjected (acceptance criteria), the characteristics of the pairs of polygons used, and the positional accuracy characterization of the BCN25 by means of a distribution function of the results obtained by means of the SBOM and the DBOM.

Firstly, with regard to the degree of fulfilment of the acceptance criteria in our approach, and thanks to the constraint imposed by the matching accuracy indicator (MAV) (which allowed us to work with only 1:1 corresponding polygons pairs among all the possible correspondences), the CC specifications were met. In addition, both GDBs were independently produced, which means that the tested source (BCN25) is not derived from the reference source (MTA10), in compliance with IC requirements. In order to meet the last set of conditions (IOC), it was necessary to transform the tested GDB (BCN25) from the ETRS89 to the ED50 reference system, in order to ensure cartographic interoperability between both GDBs. This transformation was carried out following the methodology of minimum curvature surface (MCS) developed by the National Geographic Institute of Spain [43]. The results were accurate to approximately 15 cm. If one considers that the global relative accuracy of the ED50 network is 10–20 cm, the results of the transformation using MCS were below the quality threshold of the network. Therefore, after MCS transformation the positional differences between the two GDBs due to the datum shift had no significance compared with the planimetric accuracy of the BCN (tested source), and the BCN25 and the MTA10 were finally interoperable from an cartographic point of view [19]. On the other hand, since we worked with the same type of spatial feature and representation model in both data sets, both semantic interoperability at the highest level (schema interoperability) and semantic interoperability at the object level (feature interoperability) were guaranteed.

Secondly, among all pairs of polygons obtained we used only all those pairs matched with an MAV higher or better than 0.8 (we must note that this indicator ranges between 0 and 1—for further details, see [19]). As mentioned in Section 1, the choice of this threshold value was made in order to avoid the acceptance of both erroneously-matched polygons (false positive or error of commission), which appear in the cases of 1:n or n:m correspondences, and unpaired polygons (1:0 correspondences), and was computed by assigning a confusion matrix for each BDG matching procedure. Following the results of this process, the fixed threshold guaranteed the absence of these types of errors at the 95% confidence level. The principal characteristics of both datasets and the selected pairs of polygons (MTA10 buildings and BCN25 buildings) are summarized in Table 2.

Table 2. Principal characteristics of the geospatial databases (GDBs) with regard to the buildings.

Characteristics	BCN25	MTA10
Number of polygons (total number of cases)	8863	9067
Number of polygons matched	8676	8676
Number of polygons matched with an MAV > 0.8	3356	3356
Total length of the polygons' perimeter	455,735 m	465,954 m
Mean value of the polygons' perimeter	135 m	138 m

Finally, and with regard to the characterization of the BCN25's positional accuracy, Figure 5a presents the aggregated distribution functions obtained by applying the SBOM to GDB, using buffer w with widths from 1 to 20 m. Specifically, 1 m was the size of the first buffer w_o, and the values by which it was increased (Δw, step size) were 0.1 m (for values of w between 1 and 3 m), 0.2 m (for values of w between 3 and 5 m), 0.5 m (for values of w between 5 and 10 m), and 1 m (for values of w between 10 and 20 m). The aggregated curves obtained with this method show a distribution function of the uncertainty of the BCN25 for several levels of confidence. These distributions were computed by means of a specific software tool called Matching Viewer v2016 [44]. Figure 5a shows values of around 10 m for a 95% level of confidence. On the other hand, Figure 5b shows the evolution of the distance estimated by means of Equation (1) (DBOM). This distance (6.5 m) can be considered stabilized from a buffer width of 7 m.

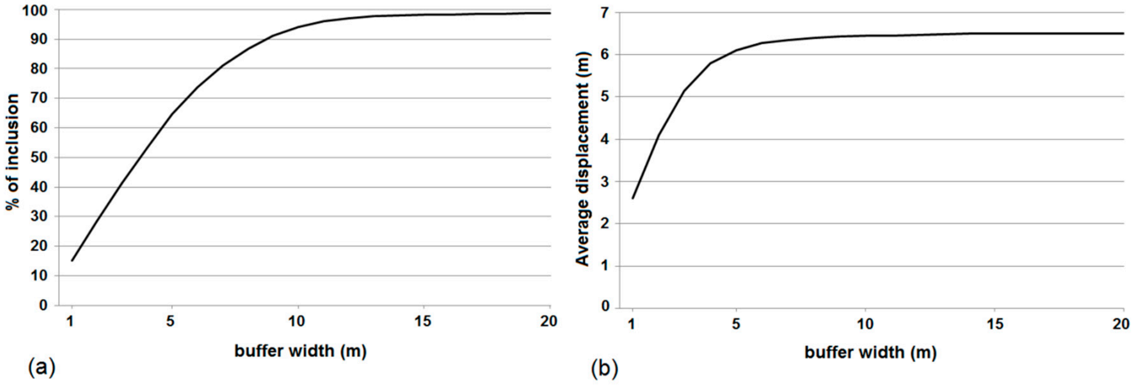

Figure 5. Aggregated distribution functions: (**a**) SBOM and (**b**) DBOM.

4. Method

This section covers the two main methodological aspects of our research: how samples of different size were obtained from the initial population, and the statistical basis of the comparison between the estimations and population values. We must note that the initial population included the subset of pairs of polygons matched with an MAV higher or better than 0.8. In addition, the parameter used for determining the different sample sizes was the length of the polygons' perimeter, measured on the polygons from the reference BDG, which was the MTA10. The reason of this last choice (in comparison with other sampling strategies based on the number of individuals) is that buffer methods are based on perimeter lines, and these, in turn, are characterized by their length. Therefore, the length of the polygons' perimeter is the most representative variable of the SBOM and DBOM methods.

4.1. Simulation Process

In order to extract samples from the initial population, a simulation process was used. The main purpose of this process is to help us understand the relationship between estimated and actual values depending on sample size, in order to obtain empirical knowledge about the sample size to use when assessing positional accuracy by means of the automatic procedure described in Section 1. Simulation can be defined as the construction of a mathematical model capable of reproducing the characteristics of a phenomenon, system, or process, in order to obtain information or solve problems [45]. Specifically, for this process we applied the Monte Carlo method [46], which requires a large amount of random executions. Therefore, our approach reproduces the APAA procedure applied to synthetic samples of polygons generated by means of a simple random sampling. Thus, the variability of the estimated planimetric accuracy of the tested GDB is obtained when applying the SBOM and the DBOM to different sample sizes.

In addition, the great advantage of using polygons (buildings) as control elements is that compared to other spatial features (lines which represent roads or coastlines), their spatial distribution may be easily controlled. In this way, the spatial distribution of control elements is always adequate, and does not affect the validity of the results. In our case, and with regard to the sampling procedure, samples (pairs of buildings represented by pairs of homologous polygons) were randomly collected from among both the urban areas and scattered rural areas that comprise our initial population. On the other hand, the number of pairs of polygons that comprise each of the samples depends on the total length L, with L being the result of adding up the individual perimeter of each polygon belonging to the MTA10. We must bear in mind that the lengths of the perimeters of two homologous polygons (each of them extracted from a different source, the MTA10 and the BCN25, respectively) are similar but not equal.

The simulation procedure was supported by the software tool called Matching Viewer v2016 [44], and consisted of the simulation of samples of different size L. These samples of different sizes (L = 5, 10, 15 . . . 100 km, where the step is ΔL = 5 km and L_o = 5) were randomly extracted from our initial population (pairs of polygons matched with an MAV higher or better than 0.8) in order to then apply them to both the SBOM and the DBOM. Specifically, for each sample size L, m samples were extracted from the initial population. Because the process is iterated m times, mean and deviation values for each parameter of interest can be computed. This process belongs to the statistical resampling technique known as bootstrap [47]. The parameters L_o and ΔL were adjusted, taking into account that 5 km represents approximately 1% (4.6 km) of the total length of perimeters.

The detailed process and its parameters are as follows:

1. An initial sample size $L = L_o$ = 5 km is considered.
2. A simple random sampling is applied to the initial population, in order to extract a sample of size L. Here the individual perimeters of the polygons belonging to the MTA10 are added up. When the sum of the perimeter lengths exceeds the length L, the last pair of polygons included in the sample that cause this excess of length is not considered. In this sense, we must note that

their exclusion had a minimal impact both on the sample size and the final results, since the mean value of the polygons' perimeter belonging to MTA10 is 138 m (Table 2), representing between 2.75% and 0.14% of L when L ranges from 5 to 100 km.

3. For both the SBOM and the DBOM, the observed distribution function (ODF) is obtained from each sample m of length L. The ODF (L,m) of the sample is compared to the population distribution function (PDF). Then, by means of the Kolmogorov–Smirnov test, the f-value and p-value are derived for each comparison case (see below).
4. Steps 2 and 3 are repeated (m = 1000 times).
5. For each sample size L, the mean values of the m iterations are derived for the f- and p-values.
6. Increase L by the step ΔL = 5 km, and repeat steps 2–5 until $L = L_{max}$ = 100 km.

4.2. Comparisons

After completing the simulation process, the comparisons between the estimated values and population values were carried out. Specifically, as stated in step number 3, the similarity between the two distribution functions (in our case, between PDF and ODF) was addressed by means of statistical tests of significance, like the Kolmogorov–Smirnov test [48,49]. Following these last authors, the results obtained by applying this test can be expressed by means of two statistic indicators: an f-value and a p-value. The first represents the maximum distance between two distribution functions and ranges in the interval [0, 1]. Thus, f-values that are close to the unit represent large discrepancies between distribution functions, while f-values close to 0 imply small discrepancies between distribution functions. On the other hand, p-values are closely linked to f-values, because they are a probabilistic measure of them. Thus, a p-value close to the unit means a great level of confidence on the corresponding f-value. For instance, when applying the Kolmogorov–Smirnov test to two distribution functions with a great similarity between them and with a high probability meeting this criterion, the f-value and p-value obtained for the pair may be 0.1 and 0.95.

Finally, we must note that we have followed the procedure described by Gibbons and Chakraborti [50] in order to develop the statistical calculations, and the p-value was approximated numerically using the method outlined by Press et al. [51].

The flowchart of the proposed method is shown in Figure 6.

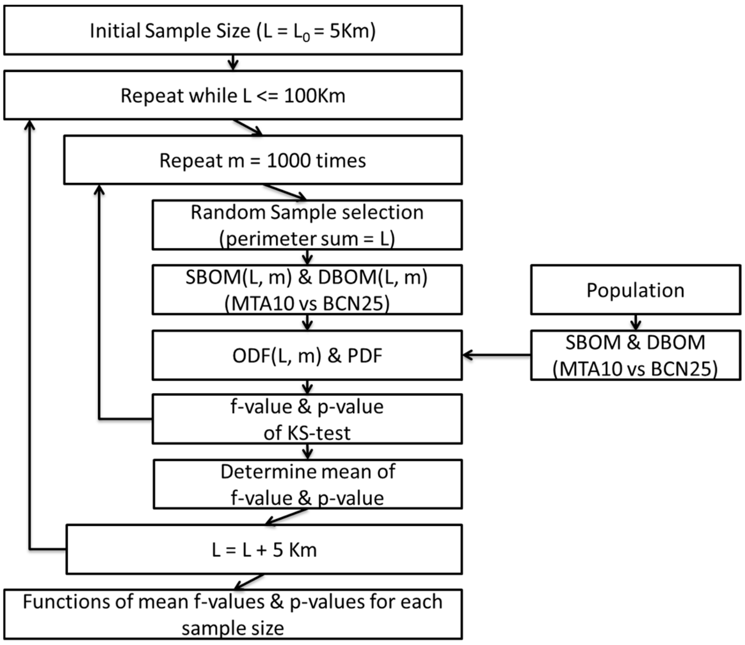
Initial Sample Size (L = L$_0$ = 5Km)
Repeat while L <= 100Km
Repeat m = 1000 times
Random Sample selection (perimeter sum = L)
SBOM(L, m) & DBOM(L, m) (MTA10 vs BCN25)
ODF(L, m) & PDF
f-value & p-value of KS-test
Determine mean of f-value & p-value
L = L + 5 Km
Functions of mean f-values & p-values for each sample size
Population
SBOM & DBOM (MTA10 vs BCN25)

Figure 6. Flowchart of the proposed method.

5. Results

The results, by which we analyze the similitude between the distribution functions (PDF and ODF), are presented by means of two types of graphical representations where the horizontal axis represents sample size (km) and the vertical axis a probability value. The first type presents the results for the frequency distance f-value (Figure 7a,b) and the second one presents the results for the associated p-value (Figure 7c,d). In addition, three different curves are represented in each graph: one of them corresponds to the mean value (represented by means of a continuous line); and the other two correspond to the 5% and 95% percentiles (represented by means of dashed lines). Finally, these graphic representations are used both for the SBOM (Figure 7a,c) and for the DBOM (Figure 7b,d).

Regarding the behaviour of the curves obtained, the first and more straightforward feature observed is that both mean f-value curves and mean p-value curves are different. In the case of mean f-value curves this difference is due to the fact that the signature given by the ODF performs in a different manner for the SBOM and for the DBOM. Obviously, the difference between mean p-value curves is due to the fact that f-values are different. In any case, these differences show that for a given sample size L, the SBOM gives better estimations than the DBOM. With regard to the shape and positioning, the 5% and 95% percentile curves are not equidistant to mean values curves (both for f-values and for p-values). In the case of f-values this means that values greater than the mean have more dispersion, while the opposite happens in the case of p-values. In addition, mean f-values, mean p-values and their associated percentile curves show a behaviour which is coherent with the supposed behaviour of an estimation process where sample size increases: f-values decrease when sample size L increases while p-values increase when sample size L increases. On the other hand, we must note that there are significant variations between f-values of the 5% and 95% percentiles when the sample size L is increased from 5 km to 100 km. Thus, for the SBOM the maximum deviation (L = 5 km) is 0.3382 and the minimum deviation (L = 100 km) is 0.0753, while in the case of the DBOM the maximum deviation (L = 5 km) is 0.7124 and the minimum deviation (L = 100 km) is 0.1656. Therefore, the reductions achieved are between 4.3 (DBOM) and 4.5 (SBOM) times the initial variability (maximum value) which correspond to small samples ($L = L_o$ = 5 km).

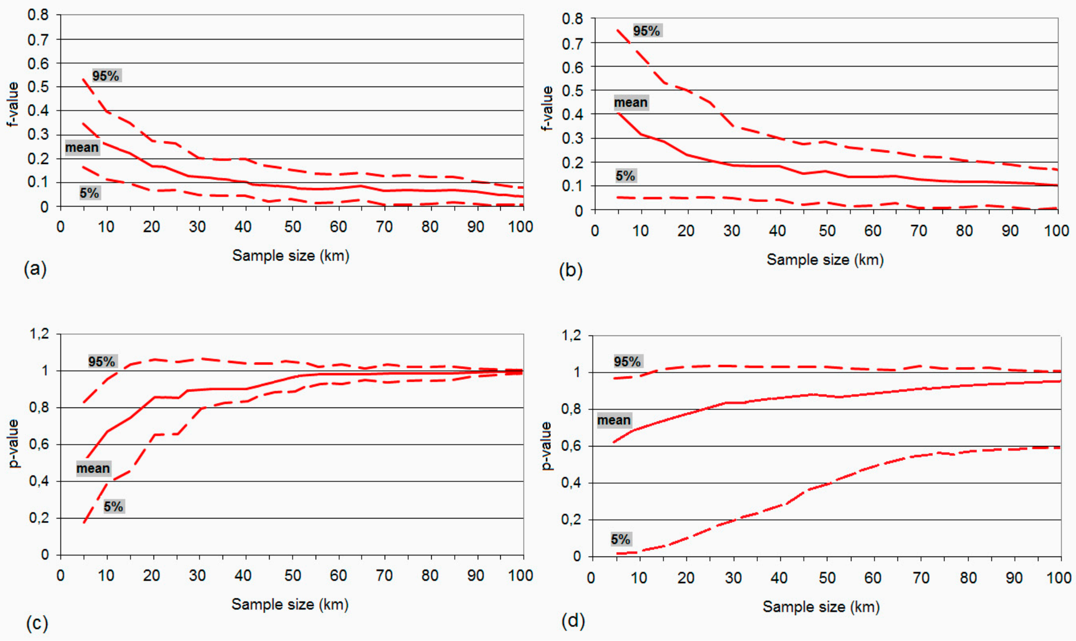

Figure 7. Statistical f-value and p-value parameters for the Kolmogorov-Smirnov test for the SBOM (**a**,**c**) and the DBOM (**b**,**d**) (1000 iterations).

Finally, and focusing our attention on the issue addressed in this paper, the curves shown in Figure 7 can be employed in order to give some guidance on the influence of sample size on APAA methods when they are used for evaluating the quality of urban GDBs. Obviously, these curves have been obtained from two specific urban GDBs (the MTA10 and the BCN25). However, they show a pattern of behavior that, in our opinion, could also be derived from other cases and scales. It will be sufficient to apply a simulation process similar to that presented here. The curves obtained can be used in two different ways:

- In order to define a sample size that will assure a certain value of mean discrepancy *f* between the sample (the ODF) and the population (the PDF);
- In order to define a sample size that will assure, with a probability of 95%, that the maximum discrepancy between the sample (the ODF) and the population (the PDF) is *f*.

Figure 8 (extracted from Figure 7) shows a practical example of the two cases described. In the first case (case 1), we wish to determine a sample size that assures a mean discrepancy of 10% between the sample and the population. In order to compute this value on the graph, we have to obtain the point where the line corresponding to level 0.1 (the *f*-value) crosses the continuous line (the mean). After obtaining this point, we have to observe the abscissa axis value (sample size) that belongs to it. In this case, *L* = 40 km (Figure 8a). In addition, the *p*-value is 90% (Figure 8b). In the second case (case 2), we wish to determine a sample size that assures, with a probability of 95%, that the maximum discrepancy between the sample (the ODF) and the population (the PDF) is 10%. Following a similar procedure to the above, we have to obtain the point where the line which corresponds to level 0.1 (the *f*-value) crosses with the dashed line (the 95% percentile value). After obtaining this point, we have to observe the abscissa axis value (sample size) that belongs to it. In this case *L* = 90 km (Figure 8a).

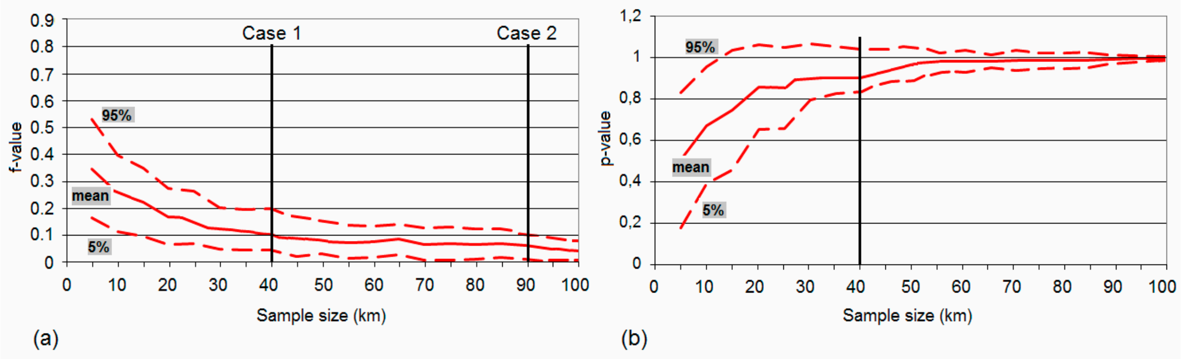

Figure 8. (**a**) Example of the use of the *f*-value graphic for determining a mean discrepancy between the sample and the population (case 1) or a maximum discrepancy (case 2); and (**b**) *p*-value obtained for case 1.

Obviously, for the case of the DBOM the procedure to follow is the same as described above. In that case, we must employ Figure 7b. Finally, taking into account the mean value of the polygons' perimeter shown in Table 2, we are able to roughly estimate the number of polygons that comprise the sample that meets the requirements outlined in the above example.

6. Conclusions

In our previous studies [19–21], we proposed an APAA methodology for GDBs, using polygonal features and two buffer-based positional accuracy assessment methods based on buffer generation on their perimeter lines: the simple buffer overlay method (SBOM) and the double buffer overlay method (DBOM). However, important aspects, such as the sample size, had not yet been adequately addressed until now.

This study addresses the influence of sample size on the variability of results derived from our APAA methodology. To that end, we employed the same two official urban GDBs used in our previous studies (the MTA10 and the BCN25), in which more than 450 km of length of perimeter (measured on 3356 pairs of polygons) were evaluated.

Our method has been based on a simulation process (supported by the software tool Matching Viewer v2016), which has consisted of the simulation of samples (randomly extracted from our initial population of polygons matched) of different size L (from 5 km to 100 km). For each sample size the simulation was iterated 1000 times. Taking into account that the results obtained by the means of the SBOM and the DBOM are expressed as distribution functions, the similarities between the various ODFs (obtained from each sample m of length L) and the PDF have been analyzed by means of the Kolmogorov–Smirnov test. The evolution of the two statistic indicators provided by this test (f-value and p-value) has allowed us to

- Gain a certain understanding of the sample size required under several different conditions concerning the mean distance value or maximum distance value between the ODFs and the PDF. This last has been confirmed by a practical example.
- Compute the variability of the estimation between the limits (sample sizes) of our simulation process. Specifically, this variability was reduced by the order of 4.5 times approximately for both methods.

Obviously, and as mentioned above, these results have been obtained from two specific urban GDBs. However, they show a pattern of behaviour that, in our opinion, could also be derived from other cases and scales, and using greater or smaller samples with greater or smaller length steps (ΔL), and running the simulation more or fewer times (m). It would be sufficient to apply a simulation process similar to that presented here.

With regard to future research, we plan to explore several directions. We plan to employ the number of polygons as the parameter used to determine the different sample sizes instead of the length of perimeter, to employ a new set of GDBs with different polygon densities, and to include new assessment methods. In addition, we will deal with the 1:n case and the n:m case, which is a multiple 1:n case. On the other hand, we are currently working on a funded research project whose aim to demonstrate the viability of our APAA approach, by means of its comparison with traditional control methods when applied to large geographical areas. As mentioned in Section 1, this viability has already been partially demonstrated for a small geographical area.

Author Contributions: The research was conducted by the second author, J.J.R.-L., under the supervision of the co-authors F.J.A.-L. and M.A.U.-C. All authors jointly drafted and critically revised the paper. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pulighe, G.; Baiocchi, V.; Lupia, F. Horizontal accuracy assessment of very high resolution Google earth images in the city of Rome, Italy. *Int. J. Digit. Earth* **2015**, *9*, 342–362. [CrossRef]
2. Goodchild, M.; Hunter, G. A Simple Positional Accuracy Measure for Linear Features. *Int. J. Geogr. Inf. Sci.* **1997**, *11*, 299–306. [CrossRef]
3. Kyriakidis, P.; Shortridge, A.; Goodchild, M. Geostatistics for conflation and accuracy assessment of digital elevation models. *Int. J. Geogr. Inf. Sci.* **1999**, *13*, 677–707. [CrossRef]
4. Song, W.; Keller, J.M.; Haithcoat, T.L.; Davis, C.H. Relaxation-based point feature matching for vector map conflation. *Trans. GIS* **2011**, *15*, 43–60. [CrossRef]
5. Smart, P.H.; Quinn, J.A.; Jones, C.B. City model enrichment. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 223–234. [CrossRef]
6. Yang, B.; Zhang, Y.; Lu, F. Geometric-based approach for integrating VGI POIs and road networks. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 126–147. [CrossRef]

7. Ruiz-Lendinez, J.J.; Ariza-López, F.J.; Ureña-Cámara, M.A.; Blázquez, E. Digital Map Conflation: A Review of the Process and a Proposal for Classification. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1439–1466. [CrossRef]
8. Lee, S.; Shan, J. Combining LIDAR elevation data and IKONOS multispectral imagery for coastal classification mapping. *Mar. Geodesy* **2003**, *26*, 117–127. [CrossRef]
9. Bartels, M.; Wei, H.; Ferryman, J. Analysis of LIDAR data fused with co-registered bands. In Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance, Sydney, Australia, 22–24 November 2006; Available online: http://www.cvg.reading.ac.uk/projects/LIDAR/index.html (accessed on 28 May 2018).
10. Chen, L.; Teo, T.; Kuo, C.; Rau, J. Shaping polyhedral buildings by the fusion of vector maps and lidar point clouds. *Photogramm. Eng. Remote Sens.* **2008**, *74*, 1147–1157. [CrossRef]
11. Elaksher, A. Fusion of hyperspectral images and lidar-based dems for coastal mapping. *Opt. Lasers Eng.* **2008**, *46*, 493–498. [CrossRef]
12. Cornet, Y.; de Béthune, S.; Binard, M.; Muller, F.; Legros, G.; Nadasdi, I. RS Data fusion using local mean and variance matching algorithms: Their respective efficiency in different urban context. In Proceedings of the IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, Rome, Italy, 8–9 November 2001; Available online: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=985698 (accessed on 28 May 2018).
13. Fanelli, A.; Leo, A.; Ferri, M. Remote sensing images data fusion: A wavelet transforms approach for urban analysis. In Proceedings of the IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, Rome, Italy, 8–9 November 2001; Available online: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=985698 (accessed on 28 May 2018).
14. Wald, L.; Ranchin, T. Data fusion for a better knowledge of urban areas. In Proceedings of the IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, Rome, Italy, 1–14 June 2001; Available online: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=985698 (accessed on 28 May 2018).
15. Xavier, E.; Ariza-López, F.J.; Ureña-Cámara, M.A. A survey of measures and methods for matching geospatial vector datasets. *ACM Comput. Surv.* **2016**, *49*, 39. [CrossRef]
16. Bruns, H.T.; Egenhofer, M.J. Similarity of spatial scenes. In Proceedings of the 7th International Symposium on Spatial Data Handling, Delft, The Netherlands, 12–16 August 1996; pp. 31–42.
17. Sheeren, D.; Mustière, S.; Zucker, J. A data-mining approach for assessing consistency between multiple representations in spatial databases. *Int. J. Geogr. Inf. Sci.* **2009**, *23*, 961–992. [CrossRef]
18. Seo, S.; O'Hara, C.G. Quality assessment of linear data. *Int. J. Geogr. Inf. Sci.* **2009**, *23*, 1503–1525. [CrossRef]
19. Ruiz-Lendínez, J.J.; Ariza-López, F.J.; Ureña-Cámara, M.A. Automatic positional accuracy assessment of geospatial databases using line-based methods. *Surv. Rev.* **2013**, *45*, 332–342. [CrossRef]
20. Ruiz-Lendínez, J.J.; Ariza-López, F.J.; Ureña-Cámara, M.A. A point-based methodology for the automatic positional accuracy assessment of geospatial databases. *Surv. Rev.* **2016**, *48*, 269–277. [CrossRef]
21. Ruiz-Lendínez, J.J.; Ureña-Cámara, M.A.; Ariza-López, F.J. A Polygon and Point-Based Approach to Matching Geospatial Features. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 399. [CrossRef]
22. Koukoletsos, T.; Haklay, M.; Ellul, C. Assessing data completeness of VGI through an automated matching procedure for linear data. *Trans. GIS* **2012**, *16*, 477–498. [CrossRef]
23. Fan, H.; Zipf, A.; Fu, Q.; Neis, P. Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 700–719. [CrossRef]
24. Ariza-López, F.J.; Xavier, E.; Ureña-Cámara, M.A. Proposal of a web service for positional quality control of spatial data sets. In Proceedings of the International Workshop on Spatial Data and Map Quality, Valletta, Malta, 20–21 January 2015.
25. Xavier, E.; Ariza-López, F.J.; Ureña-Cámara, M.A. Web service for positional quality assessment: The WPS tier. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, II-3/W5, La Grande Motte, France, 28 September–3 October 2015; pp. 257–262.
26. Xavier, E.; Ariza-López, F.J.; Ureña-Cámara, M.A. WPS for positional quality control applying the method proposed in UNE 148002. In Proceedings of the VI Jornadas Ibéricas de Infraestructuras de Datos Espaciales, Sevilla, Spain, 4–6 November 2015.
27. Herrera, F.; Lozano, M.; Verdegay, J. Tackling Real-Coded Genetic Algorithms: Operators and Tools for Behavioral Analysis. *Artif. Intell. Rev.* **1998**, *12*, 265–319. [CrossRef]

28. Tveite, H.; Langaas, S. An accuracy assessment meted for geographical line data sets based on buffering. *Int. J. Geogr. Inf. Sci.* **1999**, *13*, 27–47. [CrossRef]
29. Abbas, L.; Grussenmeyer, P.; Hottier, P. Contrôle de la planimétrie d´une base de données vectorielles: Une nouvelle méthode basée sur la distance de Hausdorff: El méthode du contrôle linéaire. *Bull. Societé Française de Photogrammétrie et Télédétection* **1995**, *137*, 6–11.
30. Kagawa, Y.; Sekimoto, Y.; Shibaski, R. Comparative Study of Positional Accuracy Evaluation of Line Data. In Proceedings of the ACRS, Hong Kong, China, 22–25 November 1999.
31. Johnston, D.; Timlin, D.; Szafoni, D.; Casanova, J.; Dilks, K. *Quality Assurance/Quality Control Procedures for ITAM GIS Databases*; US Army Corps of Engineers; Engineer Research and Development Center: Champaign, IL, USA, 2000.
32. Van Niel, T.; McVicar, T.R. Experimental evaluation of positional accuracy estimates from linear network using point and line based testing methods. *Int. J. Geogr. Inf. Sci.* **2002**, *16*, 455–473. [CrossRef]
33. Mozas-Calvache, A.T. *Exactitud Posicional de Elementos Lineales en Cartografía. Memoria Línea de Investigación Tutelada*; Departamento de Ingeniería Cartográfica, Geodésica y Fotogrametría, Universidad de Jaén: Jaén, España, 2003.
34. Hangouët, J.F. Computation of the Hausdorff distance between plane vector polylines. In Proceedings of the AutoCarto 12, Charlotte, NC, USA, 27–29 February 1995.
35. Skidmore, A.; Turner, B. Map accuracy assessment using line intersect sampling. *Photogramm. Eng. Remote Sens.* **1992**, *58*, 1453–1457.
36. Giordano, A.; Veregin, H. *Il Controllo di Qualitá nei Sistemei Informative Territoriali*; Cardo Editore: Venetia, Italia, 1994.
37. Veregin, H. Quantifying positional error induced by line simplification. *Int. J. Geogr. Inf. Sci.* **2000**, *14*, 113–130. [CrossRef]
38. Tveite, H.; Langaas, S. Accuracy assessment of geographical line data sets, the case of the digital chart of the world. In Proceedings of the 5th Scandinavian Research Conference on Geographical Information Systems (ScanGIS'95), Trondheim, Norway, 12–14 June 1995.
39. Wiedemann, C. External evaluation of road networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2003**, *34*, 93–98.
40. Chen, C.C.; Knoblock, C.A.; Shahabi, C. Automatically and accurately conflating raster maps with orthoimagery. *Geoinformatica* **2008**, *12*, 377–410. [CrossRef]
41. Cartographic Institute of Andalusia (ICA). Mapa Topográfico de Andalucía 1:10,000 Vectorial (MTA10v_2001). Available online: http://www.ideandalucia.es/catalogo/info.php (accessed on 28 May 2018).
42. National Geographic Institute of Spain (NGI). Cartographic Series: 1:25,000 and 1:50,000 Series. Available online: http://www.ign.es/ign/layout/series.do (accessed on 28 May 2018).
43. González-Matesanz, J.; Dalda, A.; Malpica, J. A Range of ED50-ETRS89 Datum Transformation Models Tested on the Spanish Geodetic Network. *Surv. Rev.* **2006**, *38*, 654–667. [CrossRef]
44. Ruiz-Lendinez, J.J.; Ureña-Cámara, M.A.; Ariza-López, F.J. Matching Viewer: Herramienta para la automatización del control de calidad posicional de la cartografía. *Rev. Colegio Oficial de Ingeniería Geomática y Topográfica* **2016**, *35*, 37–45.
45. Ríos, D.; Ríos, S.; Martín, J. *Simulación, Métodos y Aplicaciones*; Ra-Ma: Madrid, Spain, 1997.
46. Robert, C.; Casella, G. *Monte Carlo Statistical Methods*; Springer: New York, NY, USA, 2004.
47. Ross, S.M. *Simulation*, 4th ed.; Academic Press: San Diego, CA, USA, 2006.
48. William, H.; Brian, P.F.; Saul, A.T.; William, T.V. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge, UK, 1992.
49. Marsaglia, G.; Tsang, W.W.; Wang, J. Evaluating Kolmogorov's distribution. *J. Stat. Softw.* **2003**, *8*, 1–4. [CrossRef]

50. Gibbons, J.D.; Chakraborti, S. *Nonparametric Statistical Inference*, 4th ed.; Marcel Dekker, Ltd.: London, UK, 2003.
51. Press, W.A.; Vetterling, W.T.; Teukolsky, S.A.; Flannery, B.P. *Numerical Recipes in Fortran. The Art of Scientific Computing*; Cambridge University Press: Cambridge, UK, 1992.